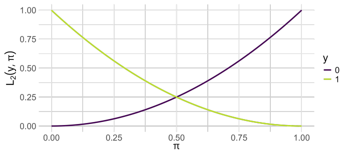
Introduction to Machine Learning

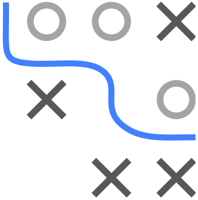
Advanced Risk Minimization
L2/L1 Loss on Probabilities



Learning goals

- Know the Brier score
- Derive the risk minimizer
- Derive the optimal constant model

BRIER SCORE



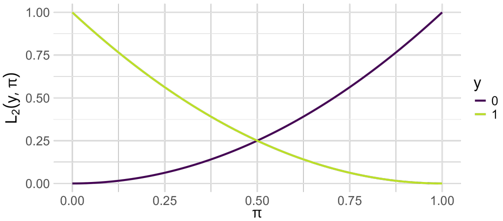
The binary Brier score is defined on probabilities $\pi \in [0, 1]$ and 0-1-encoded labels $y \in \{0, 1\}$ and is the $L2$ loss on probabilities.

$$L(y, \pi) = (\pi - y)^2$$

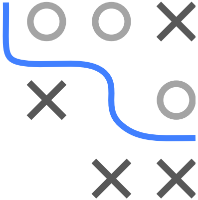
As the Brier score is a proper scoring rule (cf. section on proper scoring rules), it can be used for calibration. Despite convex in π,

$$L(y, \pi(f)) = ((1 + \exp(-f))^{-1} - y)^2$$

as a composite function is not convex in f anymore (log. sigmoid for π).



BRIER SCORE: RISK MINIMIZER



The risk minimizer for the (binary) Brier score is

$$\pi^*(\mathbf{x}) = \eta(\mathbf{x}) := \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x}),$$

which means that the Brier score attains its minimum if the prediction equals the “true” probability $\eta(\mathbf{x})$ of the outcome.

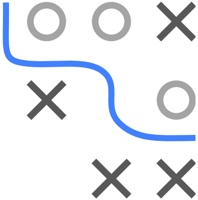
The risk minimizer for the multiclass Brier score is

$$\pi^*(\mathbf{x}) = \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}).$$

Proof: We only show the proof for the binary case. We need to minimize

$$\mathbb{E}_x \left[L(1, \pi(\mathbf{x})) \cdot \eta(\mathbf{x}) + L(0, \pi(\mathbf{x})) \cdot (1 - \eta(\mathbf{x}))\right],$$

BRIER SCORE: RISK MINIMIZER



which we do point-wise for every $\mathbf{x}$. We plug in the Brier score

$$
\begin{aligned}
& \arg\min_c \quad L(1, c)\eta(\mathbf{x}) + L(0, c)(1 - \eta(\mathbf{x})) \\
= & \arg\min_c \quad (c - 1)^2\eta(\mathbf{x}) + c^2(1 - \eta(\mathbf{x})) \quad | + \eta(\mathbf{x})^2 - \eta(\mathbf{x})^2 \\
= & \arg\min_c \quad (c^2 - 2c\eta(\mathbf{x}) + \eta(\mathbf{x})^2) - \eta(\mathbf{x})^2 + \eta(\mathbf{x}) \\
= & \arg\min_c \quad (c - \eta(\mathbf{x}))^2.
\end{aligned}
$$

The expression is minimal if $c = \eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$.

BRIER SCORE: OPTIMAL CONSTANT MODEL

The optimal constant probability model $\pi(\mathbf{x}) = \theta$ w.r.t. the Brier score for labels from $\mathcal{Y} = \{0, 1\}$ is:

$$\arg\min_{\theta} \mathcal{R}_{\text{emp}}(\theta) = \arg\min_{\theta} \sum_{i=1}^{n} \left(y^{(i)} - \theta\right)^2 = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

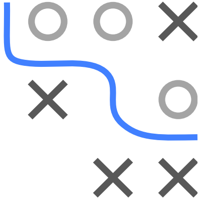
This is the fraction of class-1 observations in the observed data.
(This directly follows from our $L2$ proof for regression).

Similarly, for the multiclass brier score the optimal constant is

$$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} [y = k]$$

.

CALIBRATION AND THE BRIER SCORE



A predictor $\pi(\mathbf{x}) \in [0, 1]$ is *calibrated* if

$$\mathbb{P}\big(y = 1 \mid \pi(\mathbf{x}) = p\big) = p \quad \forall\, p \in [0, 1].$$

Intuitively, this means if we predict p, then in $100p\%$ of cases we observe $y = 1$ (neither over- or underconfident). Recall the risk minimizer for the Brier score is

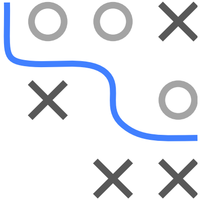
$$\pi^*(\mathbf{x}) = \eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x}).$$

Since $\pi^*(\mathbf{x}) = \eta(\mathbf{x})$ exactly, it follows that the optimal predictor satisfies

$$\mathbb{P}\big(y = 1 \mid \pi^*(\mathbf{x}) = p\big) = p,$$

i.e., is perfectly calibrated.

L1 LOSS ON PROBABILITIES



The binary L1 loss defined on probabilities $\pi \in [0, 1]$ and 0-1-encoded labels $y \in \{0, 1\}$ is given by

$$L(y, \pi) = |\pi - y|$$

As the L1 loss is not a *strictly* proper scoring rule (cf. section on proper scoring rules), it should not necessarily be used for calibration. Despite convex in π,

$$L(y, \pi(f)) = |(1 + \exp(-f))^{-1} - y|$$

as a composite function is not convex in f anymore (log. sigmoid for π).